

Weekly Report

胡万祺

一、 本周工作

【农业大数据】

1. 暑期分工

胡万祺	农业地图，将遥感影像、土壤成分等信息，在地理位置上叠加、可视化、查询。	实现 LOD 功能的农业地图，将所有已经收集到的数据可视化到对应地理位置上，并增加可视查询功能。	暑期都在
王艺	底层混合型数据库搭建，主要负责 nosql 和 redis 数据库性能测试和接口实现；实现数据分布式存储；集群数据服务器和应用服务器搭建和维护；	实现数据快速存储和实时查询；调研所有可能的农业数据格式和形式，不同的数据用不同的方法入库。	暑期都在
陈俊	底层混合型数据库搭建，主要负责 sql 数据库性能测试和接口实现；爬农业数据。	实现数据快速存储和实时查询；收集尽量多的农业数据，包括流通数据、农产品报价、土壤数据、科普知识数据等。数据获取途径：一亩田、浙江省农业信息中心、布瑞克数据等	暑期都在
张天野	围绕种植大户构建人的图谱、围绕某个农作物做农作物的知识图谱	初步构建出某个农作物(如土豆)的知识图谱；初步构建出种植大户的知识图谱	8 月 20 回实验室
王叙萌	同胡万祺的任务	同胡万祺的任务	8 月 15 日回实验室
srtp 小组	基于微软郑宇的 u-air 算法做农产品价格预测，考虑时间（历史数据）、空间临近性等因素，用滤波器去做预测。		

2. 本周工作完成情况

胡万祺：地图绘制

在网上找到比较全的 json 省市县三级地图数据，绘制出来的区域不够.shp 文件内描述的边界那么精细，但是更加美观也更加容易处理。其实我们要用到.shp 的文件的地方应该只有放缩到县级的时候。所以：国家和省的分界线可以粗略地画，可以加载一些农产品流通数

据等，突出交互方便和美观；而当用户放缩到县级地区的时候，绘制采用 gis 数据，并将 gis 数据中的维度信息如土壤、耕地空间数据等进行查看。



图 1：中国省界线绘制

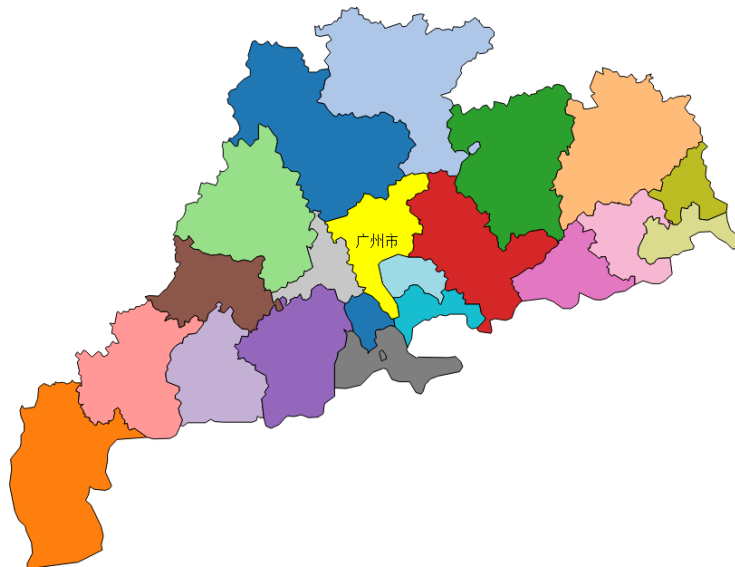


图 2：广东省市界线绘制

目前的问题：不同级别地图切换的时候放缩和投影位置没自适应，目前采用的方法是固定放缩和固定位置，先查看绘制结果；代码写的比较随意，之后可套入 angularjs 框架。

王艺：Mongodb 批量插入优化和 Restful api 设计

- (1) 修改了出租车数据的插入方式，在 Master 上插入速度为 4.2w/s，Slave 上 0.6w/s，索引 建立时间平均为 913 分钟（1.124 亿条测试数据），查询速度从 80+sd 优化至 8s（相同 条件下），还有提升空间。

详细分析：[Mongodb 批量插入与索引建立](#)

- (2) POST /DELETE/PUT/GET http://host:port/db/taxi 增删改查数据，可批量，
POST /DELETE/PUT/GET http://host:port/db/taxi/:id 增删改查 id 确定的单条数据

POST http://host:port/db/taxi/file 上传二进制数据文件做批量插入
POST http://host:port/db/taxi/zip 上传 zip 压缩的大批量数据做插入
目前的测试方式是：

1. 使用 curl 命令
2. 放在 Express 路由中，用表单提交测试

目前的问题：

1. 数据的过滤
2. 错误处理（车牌号验证、重复数据等）

详细分析：[Mongodb + Express > RESTful API 设计](#)

陈俊：PostgreSQL 插入优化查询性能测试+数据爬取

- (1) 改进了之前的 PostgreSQL 数据的插入代码，实现了 1000w 数据的插入。上个礼拜由于只测试了少量 数据，没有发现什么问题，这个礼拜测试大量数据以后发现需要做一下负载控制，辗转尝试多种方法 以后，最后决定使用 nodejs 的 async 库中的一个 cargo 函数来实现。

详见博客：[PostgreSQL 使用 Nodejs 进行海量数据批量插入](#)

- (2) 插入数据以后进行了建索引以及查询的测试，在三个列上建索引总共消费 300s。建完索引后查询测试，查询经纬度(120.83, 27.93)该点附近 1000 个点，不用索引需要 8s，使用索引则需 0.28s。

此外关于空间消费，不建索引时数据库约占 800mb，建完索引以后则约是 2.6gb。

详见博客：[Postgres 索引与 PostGIS 空间数据查询](#)

- (3) 爬取“全国农产品价格数据库”页面上的数据。进行了爬虫。主要使用 python 的 urllib 以及 BeautifulSoup 爬取该页面上的价格信息。程序目前存在以下缺陷：

- ◆ 偶尔会由于网络原因导致一些数据爬不到,但为了保证程序继续运行,不会一直重试。
- ◆ 程序偶尔会因为一些不可预测的原因中断,虽然我写了断点续爬,但我目前不确定是否会有重复的记录被爬下来。
- ◆ 程序对于爬取时间跨度大的数据(比如一次爬三个月的数据)效率较高,但如果只爬一天的数据可能导致效率十分低下。

目前已经爬取三个月的数据，文件夹结构如图 3。

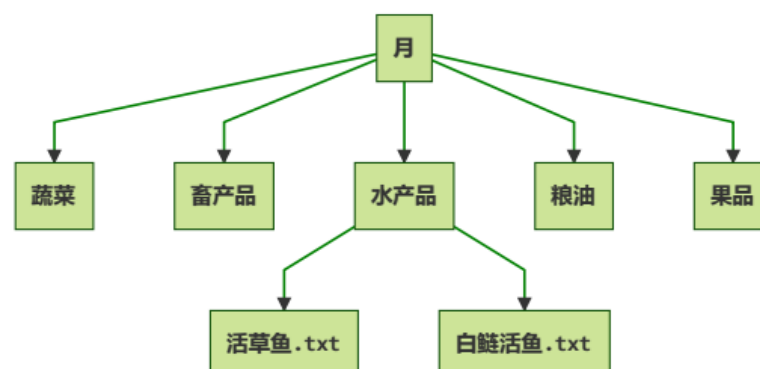


图 3：爬取数据文件夹结构

文件名实际存放会用相应的 id 而非中文字符。每条数据的格式如下：

大豆 粮油 5.2 山东省 青岛抚顺路蔬菜副食品批发市场 2015-03-31

【VisComposer】

1. 重构 UI 部分，重构逻辑如下：

UIManager

resourceswindow	资源窗口
form	
primitive	
composition	
datawindow	
workflowwindow	变换窗口
workflowui	
moduleui	
linkui	
modelwindow	
scenegraphwindow	场景图窗口
scenegraphui	
nodeui	
canvaswindow	绘制窗口

2. 完成部分重构代码：目前已完成界面基本类初始化代码，初始界面显示正常；resourcewindow 类部分代码，包括时间绑定和数据导入部分代码。目前正在整合 scenegraph 类的初始化和 workflow 的显示代码。